Express Mail Label No. **HAND DELIVERED**

UTILITY PATENT APPLICATION TRANSMITTAL
(Large Entity)

(Only for new nonprovisional applications under 37 CFR 1.53(b))

Docket No.
YOR9-2000-0324US1
Total Pages in this Submission

JC715 U.S. PTO 06/27/00

TO THE ASSISTANT COMMISSIONER FOR PATENTS
Box Patent Application
Washington, D.C. 20231

JC511 U.S. PTO 09/605709 06/27/00

Transmitted herewith for filing under 35 U.S.C. 111(a) and 37 C.F.R. 1.53(b) is a new utility patent application for an invention entitled:

AUTOMATED SET UP OF WEB-BASED CONVERSATIONAL NATURA LANGUAGE INTERFACE

and invented by:

Frederick J. Damerau and David E. Johnson

If a CONTINUATION APPLICATION, *check appropriate box and supply the requisite information:*

☐ **Continuation** ☐ **Divisional** ☐ **Continuation-in-part (CIP)** of prior application No.: _____

Which is a:

☐ **Continuation** ☐ **Divisional** ☐ **Continuation-in-part (CIP)** of prior application No.: _____

Which is a:

☐ **Continuation** ☐ **Divisional** ☐ **Continuation-in-part (CIP)** of prior application No.: _____

Enclosed are:

Application Elements

1. ☒ Filing fee as calculated and transmitted as described below
2. ☒ Specification having 9 pages and including the following:
   - a. ☒ Descriptive Title of the Invention
   - b. ☒ Cross References to Related Applications *(if applicable)*
   - c. ☐ Statement Regarding Federally-sponsored Research/Development *(if applicable)*
   - d. ☐ Reference to Microfiche Appendix *(if applicable)*
   - e. ☒ Background of the Invention
   - f. ☒ Brief Summary of the Invention
   - g. ☒ Brief Description of the Drawings *(if drawings filed)*
   - h. ☒ Detailed Description
   - i. ☒ Claim(s) as Classified Below
   - j. ☒ Abstract of the Disclosure

UTILITY PATENT APPLICATION TRANSMITTAL
(Large Entity)

(Only for new nonprovisional applications under 37 CFR 1.53(b))

Docket No. YOR9-2000-0324US1
Total Pages in this Submission

Application Elements (Continued)

3. ☒ Drawing(s) *(when necessary as prescribed by 35 USC 113)*
   a. ☐ Formal Number of Sheets _____
   b. ☒ Informal Number of Sheets 2
4. ☒ Oath or Declaration
   a. ☒ Newly executed *(original or copy)* ☐ Unexecuted
   b. ☐ Copy from a prior application (37 CFR 1.63(d)) *(for continuation/divisional application only)*
   c. ☒ With Power of Attorney ☐ Without Power of Attorney
   d. ☐ *DELETION OF INVENTOR(S)*
   Signed statement attached deleting inventor(s) named in the prior application, see 37 C.F.R. 1.63(d)(2) and 1.33(b).
5. ☐ Incorporation By Reference *(usable if Box 4b is checked)*
   The entire disclosure of the prior application, from which a copy of the oath or declaration is supplied under Box 4b, is considered as being part of the disclosure of the accompanying application and is hereby incorporated by reference therein.
6. ☐ Computer Program in Microfiche *(Appendix)*
7. ☐ Nucleotide and/or Amino Acid Sequence Submission *(if applicable, all must be included)*
   a. ☐ Paper Copy
   b. ☐ Computer Readable Copy *(identical to computer copy)*
   c. ☐ Statement Verifying Identical Paper and Computer Readable Copy

Accompanying Application Parts

8. ☒ Assignment Papers *(cover sheet & document(s))*
9. ☐ 37 CFR 3.73(B) Statement *(when there is an assignee)*
10. ☐ English Translation Document *(if applicable)*
11. ☐ Information Disclosure Statement/PTO-1449 ☐ Copies of IDS Citations
12. ☐ Preliminary Amendment
13. ☒ Acknowledgment postcard
14. ☐ Certificate of Mailing
    ☐ First Class ☐ Express Mail *(Specify Label No.):* **HAND DELIVERED**

UTILITY PATENT APPLICATION TRANSMITTAL
(Large Entity)

(Only for new nonprovisional applications under 37 CFR 1.53(b))

Docket No. YOR9-2000-0062
Total Pages in this Submission

Accompanying Application Parts (Continued)

15. ☐ Certified Copy of Priority Document(s) *(if foreign priority is claimed)*

16. ☐ Additional Enclosures *(please identify below):*

Fee Calculation and Transmittal

CLAIMS AS FILED

For	#Filed	#Allowed	#Extra		Rate	Fee
Total Claims	9	- 20 =	0	x	$18.00	$0.00
Indep. Claims	2	- 3 =	0	x	$78.00	$0.00
Multiple Dependent Claims (check if applicable) ☐						$0.00
					BASIC FEE	$690.00
OTHER FEE *(specify purpose)* _____						$0.00
					TOTAL FILING FEE	$690.00

☐ A check in the amount of to cover the filing fee is enclosed.

☒ The Commissioner is hereby authorized to charge and credit Deposit Account No. **50-0510/IBM** as described below. A duplicate copy of this sheet is enclosed.

- ☒ Charge the amount of **$690.00** as filing fee.
- ☒ Credit any overpayment.
- ☒ Charge any additional filing fees required under 37 C.F.R. 1.16 and 1.17.
- ☐ Charge the issue fee set in 37 C.F.R. 1.18 at the mailing of the Notice of Allowance, pursuant to 37 C.F.R. 1.311(b).

Signature

Dated: **June 27, 2000**

C. Lamont Whitham
Reg. No. 22,424

Whitham, Curtins & Whitham
Reston International Center
11800 Sunrise Valley Drive, Suite 900
Reston, VA 20191
(703)391-2510

cc:

LAW OFFICES
WHITHAM, CURTIS & WHITHAM, PLC
INTELLECTUAL PROPERTY LAW
11800 SUNRISE VALLEY DRIVE, SUITE 900
RESTON, VIRGINIA 20191

APPLICATION FOR UNITED STATES LETTERS PATENT

Applicants: Frederick J. Damerau and David E. Johnson
For: AUTOMATED SET UP OF WEB-BASED CONVERSATIONAL NATURAL LANGUAGE INTERFACE
Docket No.: YOR9-2000-0324US1

AUTOMATED SET UP OF WEB-BASED CONVERSATIONAL NATURAL LANGUAGE INTERFACE

CROSS-REFERENCE TO RELATED APPLICATION

This application is related to the subject matter disclosed in co-pending patent application Serial No. 09/570,788 filed May 15, 2000, by David E. Johnson, Frank J. Oles and Thilo W. Goetz for "Interactive Automated Response System" (IBM Docket YO9-99-286) and assigned to a common assignee. The disclosure of application Serial No. 09/570,788 is incorporated herein by reference.

DESCRIPTION

BACKGROUND OF THE INVENTION

Field of the Invention

The present invention generally relates to natural language systems and, more particularly, to an automated method for setting up a Web-based conversational natural language interface.

Background Description

The World Wide Web (WWW) portion of the Internet has seen an explosion of Web sites for various individual and business purposes. This in

turn has led to a growing industry in Do It Yourself (DIY) software and Web design services to assist those who want set up a Web site.

The standard method of setting up a new Web site involves a substantial amount of intellectual effort and manual processing. A typical commercial Web site might require the services of seven to nine members of a professional team working nine to fifteen months to produce. It is difficult or impossible for the average Web site administrator to do this successfully without assistance. It is even more difficult to set up a natural language query interface for a Web site once it has been built.

SUMMARY OF THE INVENTION

It is therefore an object of the present invention to provide a procedure that automates the process of setting up an instance of a conversational natural language interface for a Web site.

It is another object of the invention to automate the process of setting up a natural language interface to an existing Web site.

This invention, by automating the process of setting up a new Web site, enables a new interface to be created by anyone. Subsequent manual tuning of the interface is possible and much easier to do than creating an interface from scratch. The invention solves the problem by bringing together a number of ideas and techniques, some of which have been used in natural language processing for other purposes. In order to set up an instance of a natural language conversational interface (hereinafter NLCI), it is necessary to

(1) define a hierarchy of topics into which individual documents or Web pages can be classified,

(2) provide a keyword index for those documents for an associated search engine, and

(3) for each node in the hierarchy, specify a mechanism for associating an input natural language (NL) query to the node. (In the preferred embodiment, this mechanism is a rule set and associated rule applier.)

To solve step (1), we note that the uniform resource locators (URLs) of the Web pages associated with a single site are often organized into a coherent hierarchy of topics. On reflection, this is not surprising, since good Web design encourages logical movement from page to page. Thus, a bank might have a Web page with the URL www.bank.com/loans. It will have links to pages with URLs www.bank.com/loans/auto and www.bank.com/loans/homemortgage, and so forth. This is clearly a topic hierarchy of exactly the kind necessary for establishing the NLCI, in which "loans" is a high level node and "auto" and "homemortgage" are nodes subordinate to it. If these are the lowest level in the hierarchy, the Web pages they point to are leaves.

To solve step (2), we use methods from statistical natural language processing. From each document, we generate a set of single words, bi-grams, etc., up to n-grams for some *n*. However, these are not necessarily sequential n-grams. We allow gaps between the words making up the n-gram. The gaps are limited by establishing a distance *d* which is the maximum separation between the first and last words of the n-gram. This tactic is partial compensation for the variability allowed by natural language in expressing phrases. For example, one can say "input documents", or one might say "input text documents". The method described would generate an n-gram "input documents" from both of these. (In the preferred embodiment, words are reduced to stems, so the actual n-gram generated would be "input document".) The most frequent n-grams occurring in a document, up to some number *m*, are used as the keyword index for the document.

An example of another use for sparse n-grams, in this case bi-grams which are called "cooccurrence pairs" is explained by Ido Dagan, Shaul Marcus and Shaul Markovitch in "Contextual Word Similarity and Estimation from Sparse Data", *Association for Computational Linguistics*, pp. 164–171 (1993). The extension from bi-grams to n-grams is an obvious one.

BRIEF DESCRIPTION OF THE DRAWINGS

The foregoing and other objects, aspects and advantages will be better understood from the following detailed description of a preferred embodiment of the invention with reference to the drawings, in which:

Figure 1 is a flow diagram of the automated set up procedure according to the invention; and

Figure 2 is a block diagram showing the components of the system and their inter-relationships.

DETAILED DESCRIPTION OF PREFERRED EMBODIMENTS OF THE INVENTION

Referring now to the drawings, and more particularly to Figure 1, there is shown a flow diagram of the automated set up procedure. A program implementing a Web crawler is invoked in function block 11, beginning at the home page of the site for which a natural language interface is to be generated. The output of this module is a file of Web pages in HyperText Markup Language (HTML). In function block 12, the Uniform Resource Locators (URLs) of the Web pages are processed to induce a hierarchy of topics for the site and the HTML formatted pages are converted to the appropriate standard format. In a preferred implementation of the invention, the standard format is

eXtended Markup Language (XML). In function block 13, sparse n-grams are extracted from each page to serve as index terms for the page. The index terms are used to set up an answer generator (search engine) for the page in function block 14. In function block 15, a set of sparse n-grams is generated for each of the topics found in function block 12 by grouping together all the documents having that topic. Those n-grams satisfying some criterion for significant association with the topic are saved. In a preferred implementation of the invention, the criterion used is the chi-square measure. The sparse n-grams are converted to rules in which each term of the n-gram is a term in the rule, and the topic is the rule consequent, in function block 16. Optionally, another statistical test can be made to associate a confidence measure with each rule. In the preferred implementation of the invention, the confidence measure is the percentage of time the underlying n-gram occurs in the topic. Once the preceding steps have been accomplished, all the necessary data is at hand to finish setting up the natural language interface in function block 17. Setting up the dialog manager is accomplished according to the process described in copending patent application Serial No. 09/570,788.

Figure 2 shows the components of the system and their inter-relationships. These include the Web crawler module 21 which begins at some designated home page(s) and systematically finds all the pages reachable from these initial pages, recursively. Using the URLs of these pages, module 22 finds the topic hierarchy of this site. Note that there might be more than one root (i.e., initial home page) resulting in more than one rooted tree (hierarchy). If there is more than one rooted tree, then the final hierarchy is just

$$\begin{array}{c} \text{top} \\ \swarrow \quad \searrow \\ \text{root}_1 \quad \cdots \quad \text{root}_n \end{array}$$

with new top node "Top_n. Module 23 uses the extracted pages along with the

hierarchy to find key words and sparse phrases which can serve as index terms for the respective pages. Module 24 is an optional module for manual review and change of the decisions made by the automated system. Module 25 is a rules generating module which generates rules for each of the topics identified by module 22. Module 25 also uses the documents generated by the Web crawler module 21. The rules generated by module 25 may optionally be edited manually, as indicated by the interface between modules 24 and 25. Module 26 is the interface builder system which uses the outputs of modules 23, 25 and, optionally, 24.

While the invention has been described in terms of preferred embodiments, those skilled in the art will recognize that the invention can be practiced with modification within the spirit and scope of the appended claims.

CLAIMS

Having thus described our invention, what we claim as new and desire to secure by Letters Patent is as follows:

1. An automated method for setting up an instance of a natural language conversational interface in a Web site comprising the steps of:

defining a hierarchy of topics into which individual documents or Web pages can be classified;

generating a keyword index for those documents for an associated search engine; and

for each node in the hierarchy, specifying a mechanism for associating an input natural language (NL) query to the node.

2. The automated method for setting up an instance of a natural language conversational interface in a Web site recited in claim 1, wherein the step of generating a keyword index comprises the step of extracting sparse n-grams of keywords for each group of pages in the topic hierarchy.

3. The automated method for setting up an instance of a natural language conversational interface in a Web site recited in claim 1, further comprising the step of optionally reviewing and editing the keyword index.

4. An automated method for setting up an instance of a natural language interface in a Web site comprising the steps of:

automatically inducing a classification hierarchy by examining a structure of the Web site;

creating index terms for leaf pages from sparse n-grams; and

creating rules for a classification engine from the sparse n-grams of pages reachable from each node in a hierarchy of leaf pages, wherein each node is a classification category and the rules associated with that category are used to decide if a new input document or query reference the node.

5. The automated method for setting up an instance of a natural language interface in a Web site recited in claim 4, wherein the step of creating rules for a classification engine is performed automatically and further comprising the optional step of manually editing the rules.

AUTOMATED SET UP OF WEB-BASED CONVERSATIONAL NATURAL LANGUAGE INTERFACE

ABSTRACT OF THE DISCLOSURE

A procedure automates the process of setting up an instance of a conversational natural language interface for a Web site. By automating the process of setting up a new Web site, the process enables a new interface to be created by anyone. Subsequent manual tuning of the interface is possible and much easier to do than creating an interface from scratch. In order to set up an instance of a natural language conversational interface, it is necessary to define a hierarchy of topics into which individual documents or Web pages can be classified, provide a keyword index for those documents for an associated search engine, and for each node in the hierarchy, specify a mechanism for associating an input natural language (NL) query to the node.

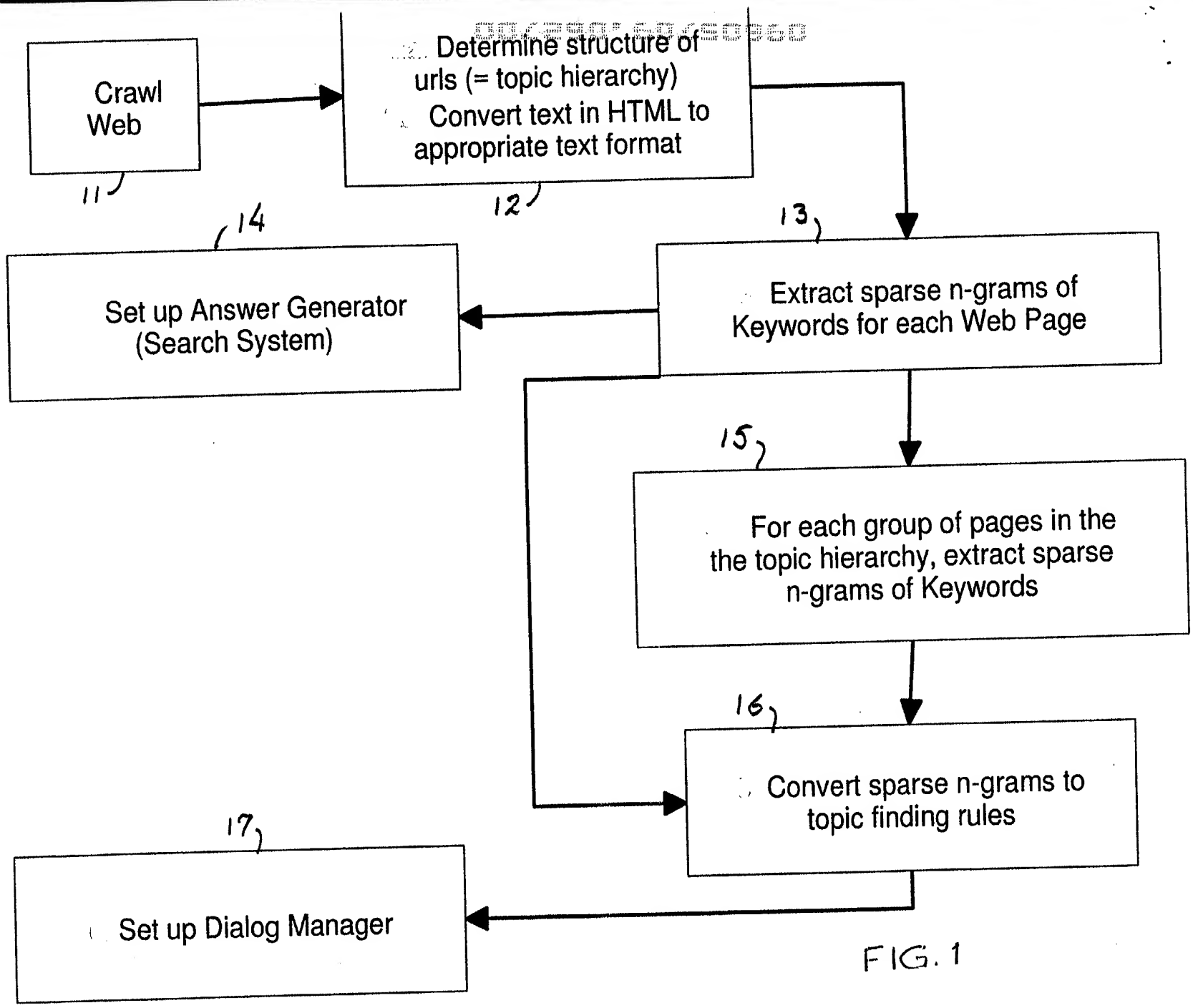
Crawl Web
11
Determine structure of urls (= topic hierarchy)
Convert text in HTML to appropriate text format
12
13
Extract sparse n-grams of Keywords for each Web Page
14
Set up Answer Generator (Search System)
15
For each group of pages in the the topic hierarchy, extract sparse n-grams of Keywords
16
Convert sparse n-grams to topic finding rules
17
Set up Dialog Manager

FIG. 1

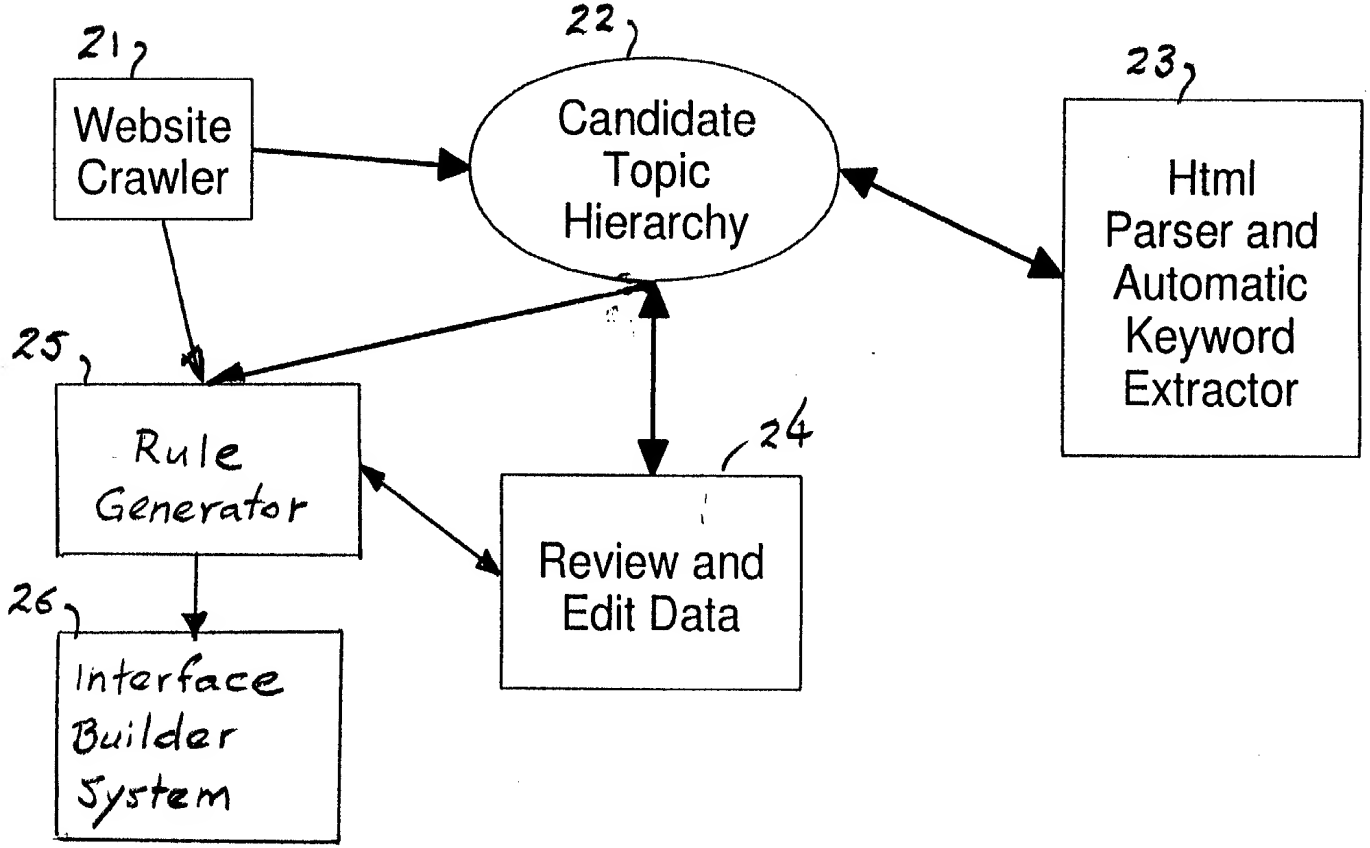
21
Website Crawler
22
Candidate Topic Hierarchy
23
Html Parser and Automatic Keyword Extractor
25
Rule Generator
24
Review and Edit Data
26
Interface Builder System

FIG. 2

Docket No.: YOR9-2000-0324US1

Application for United States Patent
Declaration and Power of Attorney

As a below named inventor, I hereby declare that:

My residence, post office address and citizenship are as stated below next to my name;

I believe I am an original, first and joint inventor of the subject matter which is claimed and for which a patent is sought on the invention entitled **AUTOMATED SET UP OF WEB-BASED CONVERSATIONAL NATURAL LANGUAGE INTERFACE** the specification of which:

(check one)

☒ is attached hereto

☐ was filed on _____ as Application Serial No. _____ and was amended on _____ (if applicable)

I hereby state that I have reviewed and understand the contents of the above identified specification, including the claims, as amended by any amendment referred to above.

I acknowledge the duty to disclose information which is material to the examination of this application in accordance with Title 37, Code of Federal Regulations, § 1.56(a).*

I hereby claim foreign priority benefits under Title 35, United States Code, §119 of any foreign application(s) for patent or inventor's certificate listed below and have also identified below any foreign application for patent or inventor's certificate having a filing date before that of the application on which priority is claimed:

Prior Foreign Application(s)

			Priority Claimed	
(Number)	(Country)	(Day/Month/Year Filed)	yes	no
(Number)	(Country)	(Day/Month/Year Filed)	yes	no

I hereby claim the benefit under Title 35, United States Code, § 120 of any United States application(s) listed below and, insofar as the subject matter of each of the claims of this application is not disclosed in the prior United States application in the manner provided by the first paragraph of Title 35, United States Code, § 112, I acknowledge the duty to disclose material information as defined in Title 37, Code of Federal Regulations, §1.56(a) which occurred between the filing date of the prior application and the national or PCT international filing date of this application:

(Application Serial No.)	(Filing Date)	(Status: patented, pending, abandoned)

Power of Attorney: As a named inventor, I hereby appoint Manny W. Schecter, Reg. No. 31,722, Terry J. Ilardi, Reg. No. 29,936, Stephen C. Kaufman, Reg. No. 29,551, Louis J. Percello, Reg. No. 33,206, Jay P. Sbrollini, Reg. No. 36,266, Robert M. Trepp, Reg. No. 25,933, Daniel P. Morris, Reg. No. 32,053, Wayne L. Ellenbogen, Reg. No. 43,602, Douglas W. Cameron, Reg. No. 31,596, David M. Shofi, Reg. No. 39,835, Christopher A. Hughes, Reg. No. 26,914, Edward A. Pennington, Reg. No. 32,588, John B. Hoel, Reg. No. 26,279, Joseph C. Redmond, Jr., Reg. No 18,753, C. Lamont Whitham, Reg. No. 22,424, Marshall M. Curtis, Reg. No. 33,138, and Michael E. Whitham, Reg. No. 32,635, as attorneys and/or agents to prosecute this application and transact all business in the Patent and Trademark Office connected therewith. All correspondence should be directed to Whitham, Curtis & Whitham, Reston International Center, 11800 Sunrise Valley Drive, Suite 900, Reston, Virginia 20191. Phone calls should be directed to Whitham, Curtis & Whitham, at 703/391-2510.

Docket No.: YOR9-2000-0324US1

I hereby declare that all statements made herein of my own knowledge are true and that all statements made on information and belief are believed to be true; and further that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code and that such willful false statements may jeopardize the validity of the application or any patent issued thereon.

(1) Inventor: Frederick J. Dameran

Signature: Frederick J Dameran Date: 6/22/2000

Residence: 356 Nash Road, North Salem, NY 10560

Citizenship: United States of America

Post Office Address: Same as Residence

(2) Inventor: David E. Johnson

Signature: David E Johnson Date: 6/23/2000

Residence: 187 Frederick Street, Cortlandt Manor, NY 10567

Citizenship: United States of America

Post Office Address: Same as Residence

*Title 37, Code of Federal Regulations, §1.56(a):
(a) A duty of candor and good faith toward the Patent and Trademark Office rests on the inventor, on each attorney or agent who prepares or prosecutes the application and on every other individual who is substantively involved in the preparation or prosecution of the application and who is associated with the inventor, with the assignee or with anyone to whom there is an obligation to assign the application. All such individuals have a duty to disclose to the Office information they are aware of which is material to the examination of the application. Such information is material where there is substantial likelihood that a reasonable examiner would consider it important in deciding whether to allow the application to issue as a patent. The duty is commensurate with the degree of involvement in the preparation or prosecution of the application.

(b) Under this section, information is material to patentability when it is not cumulative to information already of record or being made of record in the application, and (1) it establishes, by itself or in combination with other information, a prima facie case of unpatentability; or (2) it refutes, or is inconsistent with, a position the applicant takes in: (i) opposing an argument of unpatentability relied on by the Office, or (ii) asserting an argument of patentability.